Submitted to the Annals of Applied Statistics

SUPPLEMENTARY MATERIALS: LEVERAGING POPULATION OUTCOMES TO IMPROVE THE GENERALIZATION OF EXPERIMENTAL RESULTS: APPLICATION TO THE JTPA STUDY

BY MELODY HUANG[1], NAOKI EGAMI[2], ERIN HARTMAN[3], AND LUKE MIRATRIX[4]

[1]*University of California, Berkeley, melodyyhuang@berkeley.edu*

[2]*Columbia University, naoki.egami@columbia.edu*

[3]*University of California, Berkeley, ekhartman@berkeley.edu*

[4]*Harvard University, lmiratrix@g.harvard.edu*

SUPPLEMENTARY MATERIAL

1. Proofs and Derivations.

1.1. *Derivation of Variance Terms.* Consider a countably infinite population of $(\mathbf{X}_i, Y_i(t)) \sim F$, where $t \in \{0,1\}$, with density $dF(\mathbf{X}_i, Y_i(t))$. This is our target population. We define the sampling distribution for the experimental data to be $(\mathbf{X}_i, Y_i(t)) \sim \tilde{F}$ with density $d\tilde{F}(\mathbf{X}_i, Y_i(t))$. Because we consider settings where the selection into the experiment from the target population is biased, $F \neq \tilde{F}$. Let $\mathcal{S}$ be the set of all indices for all units sampled in the experimental sample. As we can consider the treatment and control groups to be independent samples from an infinite population, we will focus below on one potential outcome $Y_i(t)$.

We defined a relative density in equation (6) as follows.

$$\pi(\mathbf{X}_i) = \frac{d\tilde{F}(\mathbf{X}_i)}{dF(\mathbf{X}_i)}.$$

over the support of F, where $dF(\mathbf{X}_i) > 0$. The $\pi(\mathbf{X}_i)$ is our infinite analog to the sampling propensity score. It scales our distribution. We further assume that $\pi(\mathbf{X}_i) > 0$ (this is an overlap assumption, saying our realized sampling distribution is not missing parts of the underlying distribution). $\pi(\mathbf{X}_i)$ captures the relative density of our realized distribution to the real population. Smaller $\pi(\mathbf{X}_i)$ correspond to areas where there is a lot more in the target population than in our sample. Larger $\pi(\mathbf{X}_i)$ are where we are over-sampling.

We assume known weights for any unit, dependent on $\mathbf{X}_i$, with $w_i = \kappa/\pi(\mathbf{X}_i)$ (the κ is a fixed constant allowing our weights to be normalized on some arbitrary scale).

For the remainder of the Supplementary Materials, the distribution over which a quantity is computed will be denoted by subscript. For example, the expectation over the realized sampling distribution will be written as $\mathbb{E}_{\tilde{F}}(\cdot)$, while the expectation over the target population will be written as $\mathbb{E}_F(\cdot)$.

LEMMA 1 (Variance of a Hájek estimator). *Define $\hat{\mu}$ as a Hájek estimator:*

$$\hat{\mu}_t = \frac{\sum_{i \in \mathcal{S}} w_i Y_i(t)}{\sum_{i \in \mathcal{S}} w_i},$$

where consistent with before, $w_i = \kappa/\pi(\mathbf{X}_i)$, and $(\mathbf{X}_i, Y_i(t)) \sim \tilde{F}$. The approximate asymptotic variance of a Hájek estimator is:

$$\text{AVar}_{\tilde{F}}(\hat{\mu}_t) \approx \int \frac{1}{\pi(\mathbf{X}_i)^2}(Y_i(t) - \mu_t)^2 d\tilde{F}(\mathbf{X}_i, Y_i(t)),$$

where the asymptotic variance is being taken with respect to the realized sampling distribution, and $\mu_t = \mathbb{E}_F(Y_i(t))$ (i.e., the expected value of $Y_i(t)$ over the target population).

PROOF. To begin, we write the Hájek estimator as a ratio estimator of the following form:

$$\hat{\mu}_t = \frac{\sum_{i \in \mathcal{S}} w_i Y_i(t)}{\sum_{i \in \mathcal{S}} w_i} = \frac{\frac{1}{n}\sum_{i \in \mathcal{S}} w_i Y_i(t)}{\frac{1}{n}\sum_{i \in \mathcal{S}} w_i}$$

where we define n to be the sample size, i.e., $n = |\mathcal{S}|$.

We then define $\hat{A} = \frac{1}{n}\sum_{i \in \mathcal{S}} w_i Y_i(t)$ and $\hat{B} = \frac{1}{n}\sum_{i \in \mathcal{S}} w_i$ for notational simplicity. If we define $A = \mathbb{E}_{\tilde{F}}(\hat{A})$, $A = \kappa\mu_t$. Similarly, if we define $B = \mathbb{E}_{\tilde{F}}(\hat{B})$, $B = \kappa$.
To derive the variance expression, we will use the delta method below for a ratio, i.e., a function $h(a,b) = a/b$. For this ratio, we have

$$\frac{d}{da}h(a,b) = \frac{1}{b} \qquad \frac{d}{db}h(a,b) = -\frac{a}{b^2}.$$

Therefore, using the Delta Method for a ratio,

$$\begin{aligned}
\hat{\mu}_t &= \frac{\frac{1}{n}\sum_{i \in \mathcal{S}} w_i Y_i(t)}{\frac{1}{n}\sum_{i \in \mathcal{S}} w_i} \\
&= \frac{\hat{A}}{\hat{B}} \\
&\approx \frac{A}{B} + \frac{1}{B}(\hat{A} - A) - \frac{A}{B^2}(\hat{B} - B) \\
&= \frac{A}{B} - \frac{A}{B} + \frac{A}{B} + \frac{1}{B}\hat{A} - \frac{A}{B^2}\hat{B} \\
&= \mu_t + \frac{1}{\kappa}\frac{1}{n}\sum_{i \in \mathcal{S}} w_i Y_i(t) - \frac{\mu_t}{\kappa}\frac{1}{n}\sum_{i \in \mathcal{S}} w_i \\
&= \mu_t + \frac{1}{n\kappa}\sum_{i \in \mathcal{S}} w_i (Y_i(t) - \mu_t)
\end{aligned}$$

where the first and second equalities follow from the definition of $\hat{\mu}_t$ and $(\hat{A}, \hat{B})$, the third from the delta method, the fourth from simple algebra, the fifth from the definition of (A, B), and the sixth from re-arrangement of the terms.
Finally,

$$\begin{aligned}
\text{(A1)} \quad \text{Var}_{\tilde{F}}(\hat{\mu}_t) &= \text{Var}_{\tilde{F}}(\hat{\mu}_t - \mu_t) \\
&\approx \frac{1}{n^2\kappa^2} \cdot \text{Var}_{\tilde{F}}\left(\sum_{i \in \mathcal{S}} w_i (Y_i(t) - \mu_t)\right) \\
&= \frac{1}{n^2\kappa^2} n \int \frac{\kappa^2}{\pi(\mathbf{X}_i)^2}(Y_i(t) - \mu_t)^2 d\tilde{F}(\mathbf{X}_i, Y_i(t)) \\
\text{(A2)} \quad &= \frac{1}{n}\int \frac{1}{\pi(\mathbf{X}_i)^2}(Y_i(t) - \mu_t)^2 d\tilde{F}(\mathbf{X}_i, Y_i(t))
\end{aligned}$$

As such, $\text{AVar}_{\tilde{F}}(\hat{\mu}_t) = \lim_{n \to \infty} \text{Var}(\sqrt{n}\hat{\mu}_t) = \int \frac{1}{\pi(\mathbf{X}_i)^2}(Y_i(t) - \mu_t)^2 d\tilde{F}(\mathbf{X}_i, Y_i(t))$. □

LEMMA 2 (Weighted Variance). *Define the weighted variance and the weighted covariance as:*

$$\text{Var}_w(A_i) = \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i - \bar{A})^2 d\tilde{F}(\mathbf{X}_i, A_i)$$

$$\text{Cov}_w(A_i, B_i) = \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i - \bar{A})(B_i - \bar{B}) d\tilde{F}(\mathbf{X}_i, A_i, B_i)$$

Under this definition, common variance and covariance properties apply:

$$\text{Var}_w(A_i + B_i) = \text{Var}_w(A_i) + \text{Var}_w(B_i) + 2\text{Cov}_w(A_i, B_i)$$

$$\text{Cov}_w(A_i + B_i, C_i) = \text{Cov}_w(A_i, C_i) + \text{Cov}_w(B_i, C_i)$$

PROOF.

$$\begin{aligned}
\text{Var}_w(A_i + B_i) &= \int \frac{1}{\pi(\mathbf{X}_i)^2}\left(A_i + B_i - (\bar{A} + \bar{B})\right)^2 d\tilde{F}(\mathbf{X}_i, A_i, B_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}\left((A_i - \bar{A})^2 + (B_i - \bar{B})^2 + 2(A_i - \bar{A})(B_i - \bar{B})\right) d\tilde{F}(\mathbf{X}_i, A_i, B_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i - \bar{A})^2 d\tilde{F}(\mathbf{X}_i, A_i, B_i) + \int \frac{1}{\pi(\mathbf{X}_i)^2}(B_i - \bar{B})^2 d\tilde{F}(\mathbf{X}_i, A_i, B_i) + \\
&\quad 2\int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i - \bar{A})(B_i - \bar{B}) d\tilde{F}(\mathbf{X}_i, A_i, B_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i - \bar{A})^2 d\tilde{F}(\mathbf{X}_i, A_i) + \int \frac{1}{\pi(\mathbf{X}_i)^2}(B_i - \bar{B})^2 d\tilde{F}(\mathbf{X}_i, B_i) + \\
&\quad 2\int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i - \bar{A})(B_i - \bar{B}) d\tilde{F}(\mathbf{X}_i, A_i, B_i) \\
&= \text{Var}_w(A_i) + \text{Var}_w(B_i) + 2\text{Cov}_w(A_i, B_i)
\end{aligned}$$

$$\begin{aligned}
&\text{Cov}_w(A_i + B_i, C_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}\left(A_i + B_i - (\bar{A} + \bar{B})\right)\left(C_i - \bar{C}\right) d\tilde{F}(\mathbf{X}_i, A_i, B_i, C_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}\left((A_i - \bar{A})(B_i - \bar{B})\right)\left(C_i - \bar{C}\right) d\tilde{F}(\mathbf{X}_i, A_i, B_i, C_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}\left((A_i - \bar{A})(C_i - \bar{C}) + (B_i - \bar{B})(C_i - \bar{C})\right) d\tilde{F}(\mathbf{X}_i, A_i, B_i, C_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i - \bar{A})(C_i - \bar{C}) d\tilde{F}(\mathbf{X}_i, A_i, B_i, C_i) + \int \frac{1}{\pi(\mathbf{X}_i)^2}(B_i - \bar{B})(C_i - \bar{C}) d\tilde{F}(\mathbf{X}_i, A_i, B_i, C_i) \\
&= \int \frac{1}{\pi(\mathbf{X}_i)^2}(A_i - \bar{A})(C_i - \bar{C}) d\tilde{F}(\mathbf{X}_i, A_i, C_i) + \int \frac{1}{\pi(\mathbf{X}_i)^2}(B_i - \bar{B})(C_i - \bar{C}) d\tilde{F}(\mathbf{X}_i, B_i, C_i) \\
&= \text{Cov}_w(A_i, C_i) + \text{Cov}_w(B_i, C_i)
\end{aligned}$$

□

LEMMA 3 (Asymptotic Variance of a Weighted Estimator).
The asymptotic variance of a Hájek-style weighted estimator is:

$$\begin{aligned}\text{AVar}_{\tilde{F}}(\hat{\tau}_W) &= \text{AVar}_{\tilde{F}}(\hat{\mu}_1) + \text{AVar}_{\tilde{F}}(\hat{\mu}_0) \\ &\approx \frac{1}{p}\int \frac{1}{\pi(\mathbf{X}_i)^2}(Y_i(1)-\mu_1)^2 d\tilde{F}(\mathbf{X}_i, Y_i(1)) + \frac{1}{1-p}\int \frac{1}{\pi(\mathbf{X}_i)^2}(Y_i(0)-\mu_0)^2 d\tilde{F}(\mathbf{X}_i, Y_i(0)) \\ &= \frac{1}{p}\text{Var}_w(Y_i(1)) + \frac{1}{1-p}\text{Var}_w(Y_i(0)),\end{aligned}$$

where $\text{Var}_w(\cdot)$ *is defined in equation (9).* p *is the probability of treatment assignment, i.e.,* $p = \Pr_{\tilde{F}}(T_i = 1)$. $\mu_1 = \mathbb{E}_F(Y_i(1))$ *and* $\mu_0 = \mathbb{E}_F(Y_i(0))$.

PROOF. Because we are sampling from an infinite super-population, the treatment and control groups can be treated as two separate samples from the infinite super-population. We directly apply Lemma 1 to arrive at the final result.

□

LEMMA 4 (Asymptotic Variance of Weighted Least Squares Estimator).
The asymptotic variance of a weighted least squares estimator is:

$$\text{AVar}(\hat{\tau}_{wLS}) = \frac{1}{p}\text{Var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) + \frac{1}{1-p}\text{Var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*),$$

where γ_* *is the vector of true coefficients associated with the pretreatment covariates* $\tilde{\mathbf{X}}_i$ *defined as:*

$$(\tau_{wLS}, \alpha_*, \gamma_*) = \underset{\tau,\alpha,\gamma}{\text{argmin}}\, \mathbb{E}_{\tilde{F}}\left\{\hat{w}_i \left(Y_i - (\tau T_i + \alpha + \tilde{\mathbf{X}}_i^\top \gamma)\right)^2\right\} \tag{A3}$$

PROOF. To begin, analogous with Lin (2013) (Lemma 6), the weighted least squares estimator can be written as:
(A4)

$$\hat{\tau}_{wLS} = \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i}\sum_{i\in\mathcal{S}} w_i T_i (Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) - \frac{1}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}\sum_{i\in\mathcal{S}} w_i(1-T_i)(Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma})$$

Akin with Ding (2021), we define δ_X as:

$$\delta_X = \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i}\sum_{i\in\mathcal{S}} w_i T_i \tilde{\mathbf{X}}_i^\top - \frac{1}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}\sum_{i\in\mathcal{S}} w_i(1-T_i)\tilde{\mathbf{X}}_i^\top$$

δ_X represents any residual imbalance between the treatment and control groups in the weighted pre-treatment covariates. We can re-write Equation (A4) as:

$$\begin{aligned}\hat{\tau}_{wLS} =& \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i}\sum_{i\in\mathcal{S}} w_i T_i (Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) - \frac{1}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}\sum_{i\in\mathcal{S}} w_i(1-T_i)(Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) \\ =& \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i}\sum_{i\in\mathcal{S}} w_i T_i (Y_i(1) - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) - \frac{1}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}\sum_{i\in\mathcal{S}} w_i(1-T_i)(Y_i(0) - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) \\ =& \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i}\sum_{i\in\mathcal{S}} w_i T_i (Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_* + \tilde{\mathbf{X}}_i^\top \gamma_* - \tilde{\mathbf{X}}_i^\top \hat{\gamma}) -\end{aligned}$$

$$\frac{1}{\sum_{i\in\mathcal{S}} w_i(1-T_i)} \sum_{i\in\mathcal{S}} w_i(1-T_i)(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_* + \tilde{\mathbf{X}}_i^\top \gamma_* - \tilde{\mathbf{X}}_i^\top \hat{\gamma})$$

$$= \frac{1}{\sum_{i\in\mathcal{S}} w_i T_i} \sum_{i\in\mathcal{S}} \left(w_i T_i (Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) + w_i T_i \tilde{\mathbf{X}}_i^\top (\gamma_* - \hat{\gamma}) \right) -$$

$$\frac{1}{\sum_{i\in\mathcal{S}} w_i(1-T_i)} \sum_{i\in\mathcal{S}} \left(w_i(1-T_i)(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*) + w_i(1-T_i)\tilde{\mathbf{X}}_i^\top (\gamma_* - \hat{\gamma}) \right)$$

$$= \underbrace{\frac{1}{\sum_{i\in\mathcal{S}} w_i T_i} \sum_{i\in\mathcal{S}} w_i T_i (Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \frac{1}{\sum_{i\in\mathcal{S}} w_i(1-T_i)} \sum_{i\in\mathcal{S}} w_i(1-T_i)(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*) +}_{:=\hat{\tau}^*_{wLS}}$$

$$\underbrace{\frac{1}{\sum_{i\in\mathcal{S}} w_i T_i} \sum_{i\in\mathcal{S}} w_i T_i \tilde{\mathbf{X}}_i^\top (\gamma_* - \hat{\gamma}) - \frac{1}{\sum_{i\in\mathcal{S}} w_i(1-T_i)} \sum_{i\in\mathcal{S}} w_i(1-T_i)\tilde{\mathbf{X}}_i^\top (\gamma_* - \hat{\gamma})}_{=\delta_X(\gamma_* - \hat{\gamma})}$$

$$= \hat{\tau}^*_{wLS} + \delta_X(\gamma_* - \hat{\gamma}),$$

where $\hat{\tau}^*_{wLS}$ represents the potential outcomes, adjusted for the pre-treatment covariates using the *true* coefficients γ_*.

Under standard regularity conditions for least squares, $\gamma_* - \hat{\gamma} = o_p(1)$ (White, 1982). Furthermore, $\sqrt{n}\delta_X = O_p(1)$:

$$\lim_{n\to\infty} \mathrm{Var}_{\tilde{F}}(\delta_X) = \lim_{n\to\infty} \left(\frac{1}{n_1}\mathrm{Var}_w(\tilde{\mathbf{X}}_i) + \frac{1}{n_0}\mathrm{Var}_w(\tilde{\mathbf{X}}_i) \right)$$
$$= \lim_{n\to\infty} \frac{1}{n} \cdot \left(\frac{1}{p} + \frac{1}{1-p} \right) \mathrm{Var}_w(\tilde{\mathbf{X}}_i)$$
$$= \lim_{n\to\infty} \frac{1}{n} \cdot \frac{1}{p(1-p)} \mathrm{Var}_w(\tilde{\mathbf{X}}_i)$$

Assuming $\mathrm{Var}_w(\tilde{\mathbf{X}}_i)$ is finite, $\delta_X = O_p(\sqrt{n}^{-1}) \implies \sqrt{n}\delta_X = O_p(1)$.

Therefore, as $n \to \infty$:

$$\sqrt{n}(\hat{\tau}_{wLS} - \tau) = \sqrt{n}(\hat{\tau}^*_{wLS} - \tau) + \underbrace{\sqrt{n}\delta_X(\gamma_* - \hat{\gamma})}_{\xrightarrow{p} 0}$$

$$\xrightarrow{d} N(0, \mathrm{Var}(\hat{\tau}^*_{wLS})),$$

where $\mathrm{Var}_{\tilde{F}}(\hat{\tau}^*_{wLS}) \approx \frac{1}{p}\mathrm{Var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) + \frac{1}{1-p}\mathrm{Var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*)$ (this result follows from applying Lemma 1 on the adjusted potential outcomes).

□

1.2. *Proof of Theorem 1.*

Suppose Assumption 2 holds with $\mathbf{X}_i$, the Post-Residualized Weighted Least Squares Estimator is a consistent estimator for the PATE:

$$\hat{\tau}^{res}_{wLS} \xrightarrow{p} \tau$$

PROOF. To begin, we can write $\hat{\tau}^{res}_{wLS}$ as the above estimator on the residuals of the initial population regression:

$$\begin{aligned}
\hat{\tau}^{res}_{wLS} &= \frac{1}{(\sum_{i\in\mathcal{S}} w_i T_i)}\left(\sum_{i\in\mathcal{S}} w_i T_i(\hat{e}_i - \mathbf{X}_i\hat{\gamma}^{res})\right) - \left(\frac{1}{(\sum_{i\in\mathcal{S}} w_i(1-T_i))}\sum_{i\in\mathcal{S}} w_i(1-T_i)(\hat{e}_i - \mathbf{X}_i\hat{\gamma}^{res})\right)\\
&= \underbrace{\frac{\sum_{i\in\mathcal{S}} w_i T_i \hat{e}_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i(1-T_i)\hat{e}_i}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}}_{=\hat{\tau}^{res}_W} - \underbrace{\left(\frac{\sum_{i\in\mathcal{S}} w_i T_i \mathbf{X}_i\hat{\gamma}^{res}}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i(1-T_i)\mathbf{X}_i\hat{\gamma}^{res}}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}\right)}_{(*)},
\end{aligned}$$

where $\hat{\gamma}^{res}$ represents the estimated coefficients for the covariates $\mathbf{X}_i$ in the weighted regression run on the residualized outcomes $\hat{e}_i$. Note that the above represents two distinct regression steps: $\hat{e}_i$ is the result of the first population regression. $\hat{\gamma}^{res}$ is estimated for the covariates $\mathbf{X}_i$ from the second regression using the residualized sample outcomes, $\hat{e}_i$.

We begin by showing that $\hat{\tau}^{res}_W \xrightarrow{p} \tau$. We will begin by the proof by showing that $\hat{\tau}^{res}_W$ can be written as the difference between $\hat{\tau}_W$, and a weighted estimator computed over the fitted values $\hat{Y}_i$, which we will define as $\hat{\tau}_{\hat{Y}}$. Following the generalization literature, we treat the weights as known, as well as the observed sampled population:

$$\begin{aligned}
\hat{\tau}^{res}_W =& \frac{\sum_{i\in\mathcal{S}} w_i T_i \cdot \hat{e}_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i(1-T_i)\cdot\hat{e}_i}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}\\
=& \frac{\sum_{i\in\mathcal{S}} w_i T_i \cdot (Y_i - \hat{Y}_i)}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i(1-T_i)\cdot(Y_i - \hat{Y}_i)}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}\\
=& \underbrace{\frac{\sum_{i\in\mathcal{S}} w_i T_i \cdot Y_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i(1-T_i)\cdot Y_i}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}}_{=\hat{\tau}_W} -\\
& \underbrace{\left(\frac{\sum_{i\in\mathcal{S}} w_i T_i \cdot \hat{Y}_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i(1-T_i)\cdot\hat{Y}_i}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}\right)}_{=\hat{\tau}_{\hat{Y}}}\\
=& \hat{\tau}_W - \hat{\tau}_{\hat{Y}}
\end{aligned}$$

We will begin by showing that $\hat{\tau}_W \xrightarrow{p} \tau$. To begin:

$$\hat{\tau}_W = \frac{\sum_{i\in\mathcal{S}} w_i T_i \cdot Y_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i(1-T_i)\cdot Y_i}{\sum_{i\in\mathcal{S}} w_i(1-T_i)}$$

By Law of Large Numbers and the Continuous Mapping Theorem:

$$\hat{\tau}_W \xrightarrow{p} \underbrace{\frac{\mathbb{E}_{\tilde{F}}(w_i T_i Y_i)}{\mathbb{E}_{\tilde{F}}(w_i T_i)}}_{(1)} - \underbrace{\frac{\mathbb{E}_{\tilde{F}}(w_i(1-T_i)Y_i)}{\mathbb{E}_{\tilde{F}}(w_i(1-T_i))}}_{(2)}$$

We will now show that the first term (i.e., (1)) is equal to $\mathbb{E}_F(Y_i(1))$. We first evaluate the expectation in the denominator.

$$\begin{aligned}
\mathbb{E}_{\tilde{F}}(w_i T_i) &= \frac{n_1}{n}\mathbb{E}_{\tilde{F}}(w_i)\\
&= \frac{n_1}{n}\mathbb{E}_{\tilde{F}}\left(\frac{\kappa}{\pi(\mathbf{X}_i)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{n_1}{n} \cdot \kappa \int \frac{1}{\pi(\mathbf{X}_i)} d\tilde{F}(\mathbf{X}_i) \\
&= \frac{n_1}{n} \cdot \kappa \underbrace{\int \frac{1}{\pi(\mathbf{X}_i)} \pi(\mathbf{X}_i) dF(\mathbf{X}_i)}_{=1} \\
&= \frac{n_1}{n} \cdot \kappa
\end{aligned}
$$

For the numerator:

$$
\begin{aligned}
\mathbb{E}_{\tilde{F}}(w_i T_i Y_i) &= \mathbb{E}_{\tilde{F}}(w_i T_i Y_i(1)) \\
&= \frac{n_1}{n} \mathbb{E}_{\tilde{F}}(w_i Y_i(1)) \\
&= \frac{n_1}{n} \mathbb{E}_{\tilde{F}} \left(\frac{\kappa}{\pi(\mathbf{X}_i)} Y_i(1) \right) \\
&= \frac{n_1}{n} \cdot \kappa \mathbb{E}_{\tilde{F}} \left(\frac{1}{\pi(\mathbf{X}_i)} Y_i(1) \right) \\
&= \frac{n_1}{n} \cdot \kappa \int \frac{Y_i(1)}{\pi(\mathbf{X}_i)} d\tilde{F}(\mathbf{X}_i, Y_i(1)) \\
&= \frac{n_1}{n} \cdot \kappa \int \frac{Y_i(1)}{\pi(\mathbf{X}_i)} \cdot \pi(\mathbf{X}_i) dF(\mathbf{X}_i, Y_i(1)) \\
&= \frac{n_1}{n} \cdot \kappa \int Y_i(1) dF(\mathbf{X}_i, Y_i(1)) \\
&= \frac{n_1}{n} \kappa \cdot \mathbb{E}_F(Y_i(1))
\end{aligned}
$$

Therefore, re-writing (1):

$$
\begin{aligned}
\frac{\mathbb{E}_{\tilde{F}}(w_i T_i Y_i)}{\mathbb{E}_{\tilde{F}}(w_i T_i)} &= \frac{p\kappa \cdot \mathbb{E}_F(Y_i(1))}{p \cdot \kappa} \\
&= \mathbb{E}_F(Y_i(1))
\end{aligned}
$$

Similarly, we can show that the second term, $\mathbb{E}_{\tilde{F}}(w_i(1-T_i)Y_i)/\mathbb{E}_{\tilde{F}}(w_i(1-T_i))$, is equal to $\mathbb{E}_F(Y_i(0))$. Therefore:

$$
\begin{aligned}
\mathbb{E}_{\tilde{F}}(\hat{\tau}_W) &\xrightarrow{p} \mathbb{E}_F(Y_i(1)) - \mathbb{E}_F(Y_i(0)) \\
&= \tau
\end{aligned}
$$

Now we will show that $\hat{\tau}_{\hat{Y}} \xrightarrow{p} 0$. Once again, applying Law of Large Numbers and the Continuous Mapping Theorem:

$$
\begin{aligned}
\hat{\tau}_{\hat{Y}} &= \frac{\sum_{i \in \mathcal{S}} w_i T_i \hat{Y}_i}{\sum_{i \in \mathcal{S}} w_i T_i} - \frac{\sum_{i \in \mathcal{S}} w_i (1-T_i) \hat{Y}_i}{\sum_{i \in \mathcal{S}} w_i (1-T_i)} \\
&\xrightarrow{p} \frac{\mathbb{E}_{\tilde{F}}(w_i T_i \hat{Y}_i)}{\mathbb{E}_{\tilde{F}}(w_i T_i)} - \frac{\mathbb{E}_{\tilde{F}}(w_i(1-T_i)\hat{Y}_i)}{\mathbb{E}_{\tilde{F}}(w_i(1-T_i))} \\
&= \frac{p \cdot \mathbb{E}_{\tilde{F}}(w_i \hat{Y}_i)}{p \mathbb{E}_{\tilde{F}}(w_i)} - \frac{(1-p) \cdot \mathbb{E}_{\tilde{F}}(w_i \hat{Y}_i)}{(1-p)\mathbb{E}_{\tilde{F}}(w_i)}
\end{aligned}
$$

$$= \frac{\mathbb{E}_{\tilde{F}}(w_i \hat{Y}_i)}{\mathbb{E}_{\tilde{F}}(w_i)} - \frac{\mathbb{E}_{\tilde{F}}(w_i \hat{Y}_i)}{\mathbb{E}_{\tilde{F}}(w_i)}$$

$$= 0$$

where the third line follows from the fact that treatment assignment is randomized and independent of weights. Therefore, by the Continuous Mapping Theorem, $\hat{\tau}_W^{res} \xrightarrow{p} \tau$.

Now looking just at the $(*)$ term:

$$\frac{\sum_{i\in\mathcal{S}} w_i T_i \mathbf{X}_i \hat{\gamma}^{res}}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i) \mathbf{X}_i \hat{\gamma}^{res}}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} = \left(\frac{\sum_{i\in\mathcal{S}} w_i T_i \mathbf{X}_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i) \mathbf{X}_i}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \right) \hat{\gamma}^{res}$$

Under standard regularity conditions for least squares, $\hat{\gamma}^{res}$ converges to γ_*^{res}. Furthermore, using Law of Large Numbers and the Continuous Mapping Theorem:

$$\frac{\sum_{i\in\mathcal{S}} w_i T_i \mathbf{X}_i}{\sum_{i\in\mathcal{S}} w_i T_i} - \frac{\sum_{i\in\mathcal{S}} w_i (1-T_i) \mathbf{X}_i}{\sum_{i\in\mathcal{S}} w_i (1-T_i)} \xrightarrow{p} \frac{\mathbb{E}_{\tilde{F}}(w_i T_i \mathbf{X}_i)}{\mathbb{E}_{\tilde{F}}(w_i T_i)} - \frac{\mathbb{E}_{\tilde{F}}(w_i (1-T_i) \mathbf{X}_i)}{\mathbb{E}_{\tilde{F}}(w_i (1-T_i))}$$

$$= \frac{\mathbb{E}_{\tilde{F}}(w_i \mathbf{X}_i)}{\mathbb{E}_{\tilde{F}}(w_i)} - \frac{\mathbb{E}_{\tilde{F}}(w_i \mathbf{X}_i)}{\mathbb{E}_{\tilde{F}}(w_i)}$$

$$= 0$$

As such, we see that the term in $(*)$ will converge in probability to zero. Therefore, $\hat{\tau}_{wLS}^{res} \xrightarrow{p} \tau$. □

1.3. *Proof of Theorem 2.*

The difference between the asymptotic variance of $\hat{\tau}_W^{res}$ and the asymptotic variance of $\hat{\tau}_W$ is:

$$\text{AVar}_{\tilde{F}}(\hat{\tau}_W) - \text{AVar}_{\tilde{F}}(\hat{\tau}_W^{res})$$
$$= -\frac{1}{p(1-p)} \text{Var}_w(\hat{Y}_i) + \frac{2}{p} \text{Cov}_w(Y_i(1), \hat{Y}_i) + \frac{2}{1-p} \text{Cov}_w(Y_i(0), \hat{Y}_i),$$

PROOF. From Lemma 1.1, the asymptotic variance of a weighted estimator is:

$$\text{AVar}_{\tilde{F}}(\hat{\tau}_W) = \frac{1}{p} \text{Var}_w(Y_i(1)) + \frac{1}{1-p} \text{Var}_w(Y_i(0))$$

Using the residualized potential outcomes $\hat{e}_i(1)$ and $\hat{e}_i(0)$, the asymptotic variance of a weighted residualized estimator is:

$$\text{AVar}_{\tilde{F}}(\hat{\tau}_W^{res}) = \frac{1}{p} \text{Var}_w(\hat{e}_i(1)) + \frac{1}{1-p} \text{Var}_w(\hat{e}_i(0)).$$

From the definition of potential residuals, we can write the potential residuals as a function of the original outcome values and the fitted values:

$$\text{Var}_w(\hat{e}_i(0)) = \text{Var}_w(Y_i(0) - \hat{Y}_i)$$
$$= \text{Var}_w(Y_i(0)) + \text{Var}_w(\hat{Y}_i) - 2\text{Cov}_w(Y_i(0), \hat{Y}_i) \quad \text{(A5)}$$

$$\text{Var}_w(\hat{e}_i(1)) = \text{Var}_w(Y_i(1) - \hat{Y}_i)$$
$$= \text{Var}_w(Y_i(1)) + \text{Var}_w(\hat{Y}_i) - 2\text{Cov}_w(Y_i(1), \hat{Y}_i) \quad \text{(A6)}$$

Therefore, the difference in variances of our two estimators is

$$\begin{aligned}
&\text{AVar}_{\tilde{F}}(\hat{\tau}_W) - \text{AVar}_{\tilde{F}}(\hat{\tau}_W^{res}) \\
&= \left\{ \frac{1}{p}\text{Var}_w(Y_i(1)) + \frac{1}{1-p}\text{Var}_w(Y_i(0)) \right\} - \left\{ \frac{1}{p}\text{Var}_w(\hat{e}_i(1)) + \frac{1}{1-p}\frac{1}{n_0}\text{Var}_w(\hat{e}_i(0)) \right\} \\
&= \frac{1}{p} \cdot \big(\text{Var}_w(Y_i(1)) - \text{Var}_w(\hat{e}_i(1)) \big) + \frac{1}{1-p} \cdot \big(\text{Var}_w(Y_i(0)) - \text{Var}_w(\hat{e}_i(0)) \big)
\end{aligned}$$

Plugging in (A5) and (A6):

$$\begin{aligned}
&= -\frac{1}{p} \cdot \left\{ \text{Var}_w(Y_i(1)) + \text{Var}_w(\hat{Y}_i) - 2\text{Cov}_w(Y_i(1), \hat{Y}_i) - \text{Var}_w(Y_i(1) \right\} \\
&\quad - \frac{1}{1-p} \cdot \left\{ \text{Var}_w(Y_i(0)) + \text{Var}_w(\hat{Y}_i) - 2\text{Cov}_w(Y_i(0), \hat{Y}_i) - \text{Var}_w(Y_i(0) \right\} \\
&= -\frac{1}{p(1-p)} \cdot \text{Var}_w(\hat{Y}_i) + \frac{2}{p} \cdot \text{Cov}_w(Y_i(1), \hat{Y}_i) + \frac{2}{1-p} \cdot \text{Cov}_w(Y_i(0), \hat{Y}_i)
\end{aligned}$$

□

1.4. *Proof of Theorem 3.*

The difference between the asymptotic variance of $\hat{\tau}_{wLS}$ and the asymptotic variance of $\hat{\tau}_{wLS}^{res}$ is:

$$\begin{aligned}
&\text{AVar}_{\tilde{F}}(\hat{\tau}_{wLS}) - \text{AVar}_{\tilde{F}}(\hat{\tau}_{wLS}^{res}) \\
&= \frac{1}{p} \left\{ \text{Var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \text{Var}_w(Y_i(1) - \hat{g}(\mathbf{X}_i)) \right\} \\
&\quad + \frac{1}{1-p} \left\{ \text{Var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \text{Var}_w(Y_i(0) - \hat{g}(\mathbf{X}_i)) \right\} \\
&\quad + \frac{2}{p}\text{Cov}_w(\hat{e}_i(1), \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) + \frac{2}{1-p}\text{Cov}_w(\hat{e}_i(0), \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) - \frac{1}{p(1-p)}\text{Var}_w(\tilde{\mathbf{X}}_i^\top \gamma_*^{res}),
\end{aligned}$$

where γ_* and γ_*^{res} are the true coefficients associated with the pre-treatment covariates, $\widetilde{\mathbf{X}}_i$ defined in the weighted least squares regression (equation (13)) and the post-residualized weighted least squares regression (equation (14)), respectively. Formally, γ_* and γ_*^{res} are formally defined as the solution to the following optimization problems.

$$(\tau_{wLS}, \alpha_*, \gamma_*) = \underset{\tau,\alpha,\gamma}{\text{argmin}}\, \mathbb{E}_{\tilde{F}} \left\{ \hat{w}_i \left(Y_i - (\tau T_i + \alpha + \widetilde{\mathbf{X}}_i^\top \gamma) \right)^2 \right\} \tag{A7}$$

$$(\tau_{wLS}^{res}, \alpha_*^{res}, \gamma_*^{res}) = \underset{\tau,\alpha,\gamma}{\text{argmin}}\, \mathbb{E}_{\tilde{F}} \left\{ \hat{w}_i \left(\widehat{e}_i - (\tau T_i + \alpha + \widetilde{\mathbf{X}}_i^\top \gamma) \right)^2 \right\} \tag{A8}$$

PROOF.

$$\begin{aligned}
&\text{AVar}_{\tilde{F}}(\hat{\tau}_{wLS}) - \text{AVar}_{\tilde{F}}(\hat{\tau}_{wLS}^{res}) \\
&= \left\{ \frac{1}{p}\text{Var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) + \frac{1}{1-p}\text{Var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*) \right\}
\end{aligned} \tag{A9}$$

$$\text{(A10)} \qquad - \left\{ \frac{1}{p} \mathrm{Var}_w(\hat{e}_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) + \frac{1}{1-p} \mathrm{Var}_w(\hat{e}_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) \right\}$$

The adjusted residualized outcomes can be re-written as a function of the residualized outcomes and the fitted values from the regression. First, for the treatment outcomes:

$$\begin{aligned}
\mathrm{Var}_w(\hat{e}_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) &= \mathrm{Var}_w(Y_i(1) - \hat{g}(\mathbf{X}_i) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) \\
&= \mathrm{Var}_w(Y_i(1) - \hat{g}(\mathbf{X}_i)) + \mathrm{Var}_w(\tilde{\mathbf{X}}_i^\top \gamma_*^{res}) - 2\mathrm{Cov}_w(Y_i(1) - \hat{g}(\mathbf{X}_i), \tilde{\mathbf{X}}_i^\top \gamma_*^{res})
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{Var}_w(\hat{e}_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) &= \mathrm{Var}_w(Y_i(0) - \hat{g}(\mathbf{X}_i) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) \\
&= \mathrm{Var}_w(Y_i(0) - \hat{g}(\mathbf{X}_i)) + \mathrm{Var}_w(\tilde{\mathbf{X}}_i^\top \gamma_*^{res}) - 2\mathrm{Cov}_w(Y_i(0) - \hat{g}(\mathbf{X}_i), \tilde{\mathbf{X}}_i^\top \gamma_*^{res})
\end{aligned}$$

Plugging into Equation (A10):

$$\begin{aligned}
&\mathrm{AVar}_{\tilde{F}}(\hat{\tau}_{wLS}) - \mathrm{AVar}_{\tilde{F}}(\hat{\tau}_{wLS}^{res}) \\
&= \frac{1}{p} \left\{ \mathrm{Var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \mathrm{Var}_w(Y_i(1) - \hat{g}(\mathbf{X}_i)) \right\} \\
&\quad + \frac{1}{1-p} \left\{ \mathrm{Var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \mathrm{Var}_w(Y_i(0) - \hat{g}(\mathbf{X}_i)) \right\} \\
&\quad - \left\{ \frac{1}{p(1-p)} \mathrm{Var}_w(\tilde{\mathbf{X}}_i^\top \gamma_*^{res}) - \frac{2}{p} \mathrm{Cov}_w(Y_i(1) - \hat{g}(\mathbf{X}_i), \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) - \frac{2}{1-p} \mathrm{Cov}_w(Y_i(0) - \hat{g}(\mathbf{X}_i), \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) \right\} \\
&= \frac{1}{p} \left\{ \mathrm{Var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \mathrm{Var}_w(Y_i(1) - \hat{g}(\mathbf{X}_i)) \right\} + \frac{1}{1-p} \left\{ \mathrm{Var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*) - \mathrm{Var}_w(Y_i(0) - \hat{g}(\mathbf{X}_i)) \right\} \\
&\quad + \left\{ -\frac{1}{p(1-p)} \mathrm{Var}_w(\tilde{\mathbf{X}}_i^\top \gamma_*^{res}) + \frac{2}{p} \mathrm{Cov}_w(\hat{e}_i(1), \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) + \frac{2}{1-p} \mathrm{Cov}_w(\hat{e}_i(0), \tilde{\mathbf{X}}_i^\top \gamma_*^{res}) \right\}
\end{aligned}$$

□

1.5. *Proof of Corollary 1.*

The relative reduction in variance from residualizing is given by:

$$\text{Relative Reduction} := \frac{\mathrm{AVar}_{\tilde{F}}(\hat{\tau}_{wLS}) - \mathrm{AVar}_{\tilde{F}}(\hat{\tau}_{wLS}^{res})}{\mathrm{AVar}_{\tilde{F}}(\hat{\tau}_{wLS})} = R_0^2 - \frac{1}{1+f} \cdot \xi$$

PROOF. Let $C_1 = 1/p$ and $C_0 = 1/1-p$. Furthermore, let $\epsilon_i(1) := Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*$, $\epsilon_i(0) := Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*$, $\epsilon_i^{res}(1) := \hat{e}_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}$, $\epsilon_i^{res}(0) := \hat{e}_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*^{res}$. Then, we can write the variance of the weighted least squares estimator (i.e., Lemma 4) as:

$$\mathrm{Var}_{\tilde{F}}(\hat{\tau}_{wLS}) = \frac{1}{p} \mathrm{Var}_w(\epsilon_i(1)) + \frac{1}{1-p} \mathrm{Var}_w(\epsilon_i(0)),$$

and similarly, the variance of the residualized weighted least squares estimator as:

$$\mathrm{Var}_{\tilde{F}}(\hat{\tau}_{wLS}^{res}) = \frac{1}{p} \mathrm{Var}_w(\epsilon_i^{res}(1)) + \frac{1}{1-p} \mathrm{Var}_w(\epsilon_i^{res}(0)),$$

Then, we may re-write the relative reduction as follows:

$$\frac{\mathrm{AVar}_{\tilde{F}}(\hat{\tau}_{wLS}) - \mathrm{AVar}_{\tilde{F}}(\hat{\tau}_{wLS}^{res})}{\mathrm{AVar}_{\tilde{F}}(\hat{\tau}_{wLS})}$$

$$= \frac{C_1 \text{Var}_w(\epsilon_i(1)) + C_0 \text{Var}_w(\epsilon_i(0)) - (C_1 \text{Var}_w(\epsilon_i^{res}(1)) + C_0 \text{Var}_w(\epsilon_i^{res}(0)))}{C_1 \text{Var}_w(\epsilon_i(1)) + C_0 \text{Var}_w(\epsilon_i(0))}$$

$$= \frac{C_1 \text{Var}_w(\epsilon_i(1)) - C_1 \text{Var}_w(\epsilon_i^{res}(1)) + C_0 \text{Var}_w(\epsilon_i(0)) - C_0 \text{Var}_w(\epsilon_i^{res}(0)))}{C_1 \text{Var}_w(\epsilon_i(1)) + C_0 \text{Var}_w(\epsilon_i(0))}$$

Dividing the numerator and denominator by $C_1 \cdot \text{Var}(\epsilon_i(1))$, and defining $f = C_0 \text{Var}_w(\epsilon_i(0))/C_1 \text{Var}_w(\epsilon_i(1))$:

$$= \frac{1 - \text{Var}_w(\epsilon_i^{res}(1))/\text{Var}_w(\epsilon_i(1)) + f - f \cdot \text{Var}_w(\epsilon_i^{res}(0))/\text{Var}_w(\epsilon_i(0))}{1+f}$$

$$= \frac{1}{1+f}\left(R_1^2 + fR_0^2\right)$$

Using the definition of $\xi = R_0^2 - R_1^2$:

$$= \frac{1}{1+f}\left(R_0^2 - \xi + fR_0^2\right)$$

$$= R_0^2 - \frac{1}{1+f} \cdot \xi$$

□

2. Diagnostic Measure. We detail how to estimate the diagnostic measures in this section. To estimate the diagnostic for the post-residualized weighted estimator, we compute the estimated weighted variance of both the residuals and the outcomes for the units assigned to control:

$$\hat{R}_0^2 = 1 - \frac{\widehat{\text{Var}}_{w,0}(\hat{e}_i)}{\widehat{\text{Var}}_{w,0}(Y_i)}$$

$$= 1 - \frac{\sum_{i \in \mathcal{S}} w_i^2 (1-T_i)(\hat{e}_i - \hat{\mu}_0^{res})^2}{\sum_{i \in \mathcal{S}} w_i^2 (1-T_i)(Y_i - \hat{\mu}_0)^2} \tag{A11}$$

where $\hat{\mu}_0$ and $\hat{\mu}_0^{res}$ are defined as:

$$\hat{\mu}_0 = \frac{\sum_{i \in \mathcal{S}} w_i (1-T_i) Y_i}{\sum_{i \in \mathcal{S}} w_i (1-T_i)}, \qquad \hat{\mu}_0^{res} = \frac{\sum_{i \in \mathcal{S}} w_i (1-T_i) \hat{e}_i}{\sum_{i \in \mathcal{S}} w_i (1-T_i)} \tag{A12}$$

For the post-residualized weighted least squares estimator, estimating the diagnostic follows similarly, but we now have to account for the covariate adjustment taking place:

$$\hat{R}_{0,wLS}^2 = 1 - \frac{\widehat{\text{Var}}_{w,0}(\hat{e}_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}_0^{res})}{\widehat{\text{Var}}_{w,0}(Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}_0)}$$

$$= 1 - \frac{\sum_{i \in \mathcal{S}} w_i^2 (1-T_i)(\hat{\epsilon}_i^{res} - \hat{\epsilon}_0^{res})^2}{\sum_{i \in \mathcal{S}} w_i^2 (1-T_i)(\hat{\epsilon}_i - \hat{\epsilon}_0)^2}, \tag{A13}$$

where $\hat{\epsilon}_i$ represents the residuals estimated from regressing the outcomes Y_i on the pre-treatment covariates $\tilde{\mathbf{X}}_i$, across the subset of units assigned to control (i.e., $Y_i - \tilde{\mathbf{X}}_i^\top \hat{\gamma}_0$, where $\hat{\gamma}_0$ is estimated by running the regression $Y_i \sim \tilde{\mathbf{X}}_i$ across units assigned to control). $\hat{\epsilon}_i^{res}$ is analogously defined for the residualized outcomes $\hat{e}_i$. $\hat{\epsilon}_0$ and $\hat{\epsilon}_0^{res}$ are the weighted average of both $\hat{\epsilon}_i$ and $\hat{\epsilon}_i^{res}$, respectively.

When treating $\hat{Y}_i$ as a covariate, the diagnostic can be estimated in an analogous way, but by first performing sample splitting. More specifically, the procedure for including $\hat{Y}_i$ as a covariate for the weighted estimator is as follows:

1. Across the subset of units assigned to control, randomly partition the units into two subsets: S_1 and S_2. Without loss of generality, we will use S_1 as our training sample, and S_2 as our testing sample.
2. Regress $\hat{Y}_i$ on the outcomes across S_1 to obtain a $\hat{\beta}$ value.
3. Using $\hat{\beta}$, estimate the out-of-sample residuals $\hat{e}_i^{oos}$ across S_2, where $\hat{e}_i^{oos} := Y_i - \hat{\beta}\hat{Y}_i$.
4. Estimate the diagnostic using $\hat{e}_i^{oos}$ and the outcomes Y_i across S_2 using Equation (A11).
5. Cross-fit: repeat steps 1-3, but flipping S_1 and S_2 (i.e., regress $\hat{Y}_i$ on the outcomes across S_2 to obtain a $\hat{\beta}$ value, and estimate the diagnostic across S_1).
6. Average the two diagnostic values together.

When including $\hat{Y}_i$ as a covariate for the weighted least squares estimator, researchers can repeat the procedure above; however, when estimating the diagnostic using $\hat{e}_i^{oos}$, researchers must account for $\tilde{\mathbf{X}}_i$. More specifically:

1. Follow Steps 1-3 above to obtain $\hat{e}_i^{oos}$ across S_1.
2. Regress $\hat{e}_i^{oos}$ on $\tilde{\mathbf{X}}_i$, and regress Y_i on $\tilde{\mathbf{X}}_i$ across S_2. Use Equation (A13) to estimate the diagnostic value.
3. Cross fit, and average the two diagnostic values together.

When researchers have relatively small sample sizes, it can be advantageous to perform repeated sample splitting, and take the average of the diagnostic across all the repeated splits (see Jacob (2020) for more details).

3. Simulations. This section provides details associated with the simulations described in Section 6 of the main manuscript.

3.1. *Simulation Set-Up.* To begin, we randomly generate a set of covariates $\begin{bmatrix} X_1 \; X_2 \; X_S \; X_\tau \end{bmatrix} \sim MVN(\mathbf{0}, \Sigma)$ with the following covariance structure:

$$\Sigma = \begin{bmatrix} 1 & 0 & 0.45 & 0.5 \\ 0 & 1 & 0 & 0 \\ 0.45 & 0 & 1 & 0.9 \\ 0.5 & 0 & 0.9 & 1 \end{bmatrix}$$

where, recall, (X_{1i}, X_{2i}) are observed pre-treatment covariates, X_{Si} controls the probability of inclusion in the experimental sample, and $X_{\tau i}$ determines the treatment effect.

Unit i's propensity for being included in the experimental sample (recorded as $S_i = 1$) is governed by a logit model on the covariate X_{Si}:

$$P(S_i = 1) \propto \frac{\exp(X_{Si})}{1 + \exp(X_{Si})}.$$

At each iteration of the simulation, an experimental sample is drawn using the propensity score, as well as a random sample of the population. The sampled population is used to estimate the residualizing model and sampling weights.

Each specific data generating process for the potential outcome under control is determined by the values of the βs and γs and α. Below, we provide the parameter values and simplified DGP for $Y_i(0)$.

- Scenario 1: Linear Data Generating Process, identical population/sample DGP
  $\beta_1 = 2, \beta_2 = 1, \beta_3 = 0, \beta_S = 0, \gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0, \gamma_4 = 0, \alpha = 0$, yielding:

$$Y_i(0) = 2X_{1i} + X_{2i} + \varepsilon_i$$

Summary of Estimator Performance (N=10,000)

		DiM	Weighted $\widehat{\tau}_W$	Weighted $\widehat{\tau}_W^{res}$	Weighted $\widehat{\tau}_W^{cov}$	Weighted Least Squares $\widehat{\tau}_{wLS}$	Weighted Least Squares $\widehat{\tau}_{wLS}^{res}$	Weighted Least Squares $\widehat{\tau}_{wLS}^{cov}$
Scenario 1: Linear Outcome Model								
n=100	MSE	36.44	30.05	1.48	1.34	1.34	1.34	1.30
	Bias	3.60	-0.13	0.05	0.12	0.19	0.19	0.27
	SE	4.85	5.48	1.22	1.15	1.14	1.14	1.11
n=1000	MSE	16.41	2.98	0.17	0.15	0.14	0.14	0.13
	Bias	3.74	0.00	-0.01	0.00	0.00	0.00	0.01
	SE	1.56	1.73	0.41	0.38	0.38	0.38	0.36
n=5000	MSE	14.39	0.64	0.04	0.03	0.03	0.03	0.03
	Bias	3.72	0.01	0.00	0.00	0.01	0.01	0.01
	SE	0.72	0.80	0.19	0.19	0.18	0.18	0.18
Scenario 2: Nonlinear Outcome Model								
n=100	MSE	70.71	58.80	8.25	8.20	36.59	8.16	8.04
	Bias	3.44	-0.30	0.09	0.14	0.04	0.23	0.26
	SE	7.68	7.67	2.87	2.86	6.05	2.85	2.83
n=1000	MSE	20.37	5.58	0.82	0.80	3.53	0.79	0.78
	Bias	3.78	0.05	-0.00	-0.00	0.05	0.00	0.01
	SE	2.46	2.36	0.91	0.90	1.88	0.89	0.89
n=5000	MSE	14.80	1.17	0.18	0.18	0.83	0.17	0.17
	Bias	3.68	-0.02	-0.01	-0.01	-0.03	-0.01	-0.00
	SE	1.12	1.08	0.42	0.42	0.91	0.42	0.42

TABLE A1

Summary of estimator performance for Scenarios 1 and 2. The population is fixed at $N = 10,000$, and $1,000$ iterations were run for each sample size. MSE is scaled by 100, and the bias and standard error are scaled by 10.

- Scenario 2: Nonlinear Data Generating Process, identical population/sample DGP
  $\beta_1 = 2, \beta_2 = 1, \beta_3 = 0, \beta_S = 2.5, \gamma_1 = 0.5, \gamma_2 = 3, \gamma_3 = 2.5, \gamma_4 = 0, \alpha = 0$, yielding:
  $$Y_i(0) = 2X_{1i} + X_{2i} + 0.5X_{1i}^2 + 3\sqrt{|X_{2i}|} + 2.5\big(X_{1i} \cdot X_{2i}\big) + \varepsilon_i$$
- Scenario 3: Linear Data Generating Process, different population/sample DGP
  $\beta_1 = 2, \beta_2 = 1, \beta_3 = -1, \beta_S = \beta_S, \gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0, \gamma_4 = 0, \alpha = 0.5$, yielding:
  $$Y_i(0) = 2X_{1i} + X_{2i} + \beta_S \cdot (1 - S_i) \cdot (0.5 - X_{1i}) + \varepsilon_i,$$
- Scenario 4: Nonlinear Data Generating Process, different population/sample DGP
  $\beta_1 = 2, \beta_2 = 1, \beta_3 = -1, \beta_S = \beta_S, \gamma_1 = 0.5, \gamma_2 = 3, \gamma_3 = 2.5, \gamma_4 = 1.5, \alpha = 0.5$, yielding:
  $$\begin{aligned} Y_i(0) =& 2X_{1i} + X_{2i} + 0.5X_{1i}^2 + 3\sqrt{|X_{2i}|} + 2.5\big(X_{1i} \cdot X_{2i}\big) \\ & \beta_S \cdot (1 - S_i) \cdot (0.5 - X_{1i} + 1.5X_{1i} \cdot X_{2i}) + \varepsilon_i, \end{aligned}$$

For Scenarios 3 and 4, β_S takes on values $\{-5, -2, -1, 0, 1, 2, 5\}$.

3.2. *Supplementary Tables.* Table A1 presents summary results for estimator performance under Scenarios 1 and 2, including MSE, Bias, and SE. Column 1 presents the baseline results for the difference-in-means (DiM). Columns 2-4 present the results for the weighted estimators and columns 5-7 present results for the weighted least squares estimator. For the weighted and weighted least squares estimators we present the standard estimator without residualizing, the directly residualized estimator and inclusion of $\hat{Y}$ as a covariate.

Table A2 presents summary results for estimator performance under Scenarios 3 and 4, including MSE and Bias. In these scenarios we vary the value of β_S, presented in column 1, which controls the degree of alignment between the experimental sample outcomes and

Summary of Estimator Performance - Scenario 3 and 4 (N = 10,000)

	DiM		Weighted						Weighted Least Squares					
			$\widehat{\tau}_W$		$\widehat{\tau}_W^{res}$		$\widehat{\tau}_W^{cov}$		$\widehat{\tau}_{wLS}$		$\widehat{\tau}_{wLS}^{res}$		$\widehat{\tau}_{wLS}^{cov}$	
β_S	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias
Scenario 3: Linear Outcome														
-5	16.41	3.74	2.98	0.00	10.69	-0.11	0.36	-0.03	0.14	0.00	0.14	0.00	0.13	0.01
-2.5	15.83	3.67	3.07	-0.06	2.55	0.06	0.25	0.01	0.14	0.02	0.14	0.02	0.13	0.04
-2	16.05	3.72	2.99	0.01	1.54	0.02	0.22	0.02	0.14	0.02	0.14	0.02	0.14	0.04
-1	16.11	3.73	2.88	0.05	0.39	-0.02	0.16	0.00	0.14	-0.00	0.14	-0.00	0.13	0.02
-0.5	16.37	3.75	2.89	0.07	0.17	-0.02	0.14	-0.00	0.13	0.00	0.13	0.00	0.13	0.02
0	16.50	3.75	3.04	0.06	0.16	0.00	0.14	0.00	0.13	0.00	0.13	0.00	0.13	0.02
0.5	16.38	3.74	3.19	0.04	0.41	0.01	0.21	0.01	0.13	0.01	0.13	0.01	0.12	0.02
1	16.11	3.72	3.03	0.00	0.92	0.01	0.54	0.02	0.13	-0.01	0.13	-0.01	0.12	0.01
2	16.23	3.74	3.03	0.01	2.68	0.04	2.68	0.05	0.14	-0.00	0.14	-0.00	0.13	0.01
2.5	16.09	3.71	3.15	-0.01	3.92	0.01	3.15	-0.00	0.14	-0.01	0.14	-0.01	0.13	0.01
5	16.33	3.71	3.23	0.00	14.32	-0.01	1.54	0.02	0.14	-0.00	0.14	-0.00	0.13	0.01
Scenario 4: Nonlinear Outcome														
-5	20.31	3.74	5.77	0.04	37.03	-0.01	5.66	0.04	3.72	0.05	26.19	0.10	1.02	0.03
-2.5	20.31	3.74	6.17	-0.01	9.55	0.10	5.10	0.04	3.96	0.05	7.57	0.06	1.67	0.04
-2	19.50	3.65	5.92	-0.08	6.22	-0.00	4.27	-0.04	3.86	-0.04	5.05	-0.05	2.89	-0.00
-1	19.77	3.73	5.71	-0.02	2.18	-0.08	2.14	-0.07	3.91	-0.09	1.92	-0.05	1.08	0.02
-0.5	19.75	3.68	5.70	-0.06	1.10	-0.03	1.09	-0.03	3.96	-0.12	1.06	-0.01	0.83	0.05
0	19.74	3.69	5.81	-0.04	0.81	0.01	0.80	0.01	3.71	-0.05	0.77	0.02	0.77	0.02
0.5	20.49	3.83	5.40	0.09	1.42	0.03	1.30	0.03	3.65	0.04	1.09	0.01	0.75	0.02
1	20.24	3.80	5.52	0.08	2.84	-0.05	2.04	-0.01	3.95	0.06	1.91	-0.07	0.80	-0.01
2	20.03	3.72	5.83	0.05	7.99	-0.02	3.04	0.02	4.24	0.03	5.27	-0.06	0.84	-0.00
2.5	20.45	3.74	6.04	0.06	12.15	-0.11	3.51	-0.01	4.28	0.05	8.32	-0.09	0.85	-0.00
5	20.80	3.75	6.29	0.08	45.95	-0.25	5.05	0.02	4.09	0.06	29.97	-0.27	0.92	-0.02

TABLE A2

Summary of estimator performance for Scenarios 3 and 4, where $n = 1,000$ and $N = 10,000$. 1,000 iterations were run for each β_S value. The bias is scaled by 10, and the MSE is scaled by 100.

the population outcomes. We fix the experimental sample size at $n = 1,000$. Columns 2-3 presents the baseline results for the difference-in-means (DiM). Columns 4-9 present the results for the weighted estimators and columns 10-15 present results for the weighted least squares estimator. For the weighted and weighted least squares estimators we present the standard estimator without residualizing, the directly residualized estimator and inclusion of $\hat{Y}$ as a covariate.

In Table A3 we summarize the true positive and true negative rates for the diagnostic measures for the post-residualized estimators.[1] Column 1 presents the value of β_S. Columns 2-9 present the post-residualized weighted, post-residualized weighted least squares, the post-residualized weighted estimator with $\hat{Y}$ as a covariate, and the post-residualized weighted least squares estimator with $\hat{Y}$ as a covariate, respectively. We see that in general, the diagnostic measures are able to adequately capture when residualizing results in precision gain. We see that using sample splitting to estimate the pseudo-R^2 measure for the case in which we include $\hat{Y}_i$ as a covariate can sometimes be conservative, which results in a low true positive rate in cases when the divergence between the experimental sample and population are rather large. In cases where residualizing always leads to losses or gains in precision, the total number of true positive or true negative rates is zero (respectively).

Finally, in Table A4 we evaluate the 95% coverage rates for the proposed post-residualized estimators. We see that in all scenarios, we achieve at least nominal coverage. When the population and sample data generating processes diverge significantly, we showed in the previous

[1] True positive rates were calculated by taking the total number of true positives (i.e., cases where the diagnostic correctly indicated there would be efficiency gain from residualizing) and dividing by the total number of cases in which residualizing led to efficiency gain. True negatives are similarly defined.

Diagnostic Performance across Simulations

	$\hat{\tau}_W^{res}$		$\hat{\tau}_W^{cov}$		$\hat{\tau}_{wLS}^{res}$		$\hat{\tau}_{wLS}^{cov}$	
β_S	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
Scenario 3: Linear Outcomes								
-5	0/0	1000/1000	1000/1000	0/0	207/472	329/528	338/705	166/295
-2.5	1/942	58/58	1000/1000	0/0	203/499	304/501	308/694	177/306
-2	999/1000	0/0	1000/1000	0/0	216/514	288/486	310/689	175/311
-1	1000/1000	0/0	1000/1000	0/0	219/525	287/475	293/689	188/311
-0.5	1000/1000	0/0	1000/1000	0/0	214/519	282/481	293/689	183/311
0	1000/1000	0/0	1000/1000	0/0	223/523	275/477	283/683	177/317
0.5	1000/1000	0/0	1000/1000	0/0	222/536	268/464	260/666	194/334
1	1000/1000	0/0	1000/1000	0/0	233/519	283/481	254/669	199/331
2	999/1000	0/0	998/1000	0/0	228/490	321/510	297/695	175/305
2.5	0/0	999/1000	188/490	346/510	209/466	336/534	341/705	149/295
5	0/0	1000/1000	1000/1000	0/0	214/486	303/514	322/699	155/301
Scenario 4: Nonlinear Outcomes								
-5	0/0	1000/1000	360/718	224/282	0/0	1000/1000	58/1000	0/0
-2.5	0/0	998/1000	881/985	10/15	0/0	1000/1000	0/1000	0/0
-2	87/217	738/783	950/996	2/4	0/0	998/1000	0/994	5/6
-1	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0
-0.5	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0
0	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0
0.5	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0	1000/1000	0/0
1	1000/1000	0/0	1000/1000	0/0	999/1000	0/0	1000/1000	0/0
2	13/28	907/972	1000/1000	0/0	22/28	906/972	1000/1000	0/0
2.5	0/0	1000/1000	1000/1000	0/0	0/0	1000/1000	1000/1000	0/0
5	0/0	1000/1000	999/1000	0/0	0/0	1000/1000	1000/1000	0/0

TABLE A3
True positive rates (TPR) and true negative rates (TNR) for the diagnostic measures.

sections that there could be a loss in efficiency from using post residualized weighting. However, coverage rates are not affected by residualizing.

4. Additional Information for Empirical Application. As discussed in Section 7, we construct our target population using a leave-one-out procedure. Table A5 provides a summary of the site specific and target population average treatment effects. More specifically, the difference-in-means (DiM) columns denote the experimental estimate in the specific site. The target PATE is defined as the average difference-in-means estimate across the other 15 sites. Standard errors are presented in parentheses. Certain sites, such as MT (Butte, MT) contain only 38 experimental units, and the point estimate of the experimental site DiM is vastly different from the target PATE. Thus, we expect the task of generalizing to be more difficult for these sites.

4.1. *Estimating the Residualizing Model.* Pre-treatment covariates were taken from the baseline survey conducted at the beginning of the original JTPA experiment, to assess whether or not individuals were eligible for JTPA services. A full list of the covariates included in the residualizing model is provided in Table A6. In addition to the pre-treatment covariates, we also include normalized measures of previous earnings. Specifically, we include the z-score of an individual's previous earnings, relative to the experimental site, as well as the z-score of an individual's previous earnings, relative to the entire population.

Coverage Rates

β_S	Weighted $\widehat{\tau}_W$	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	Weighted Least Squares $\widehat{\tau}_{wLS}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$
Scenario 3: Linear Outcome						
-5	0.95	0.95	0.97	0.99	0.99	0.99
-2.5	0.95	0.96	0.97	0.98	0.98	0.98
-2	0.95	0.97	0.98	0.97	0.97	0.98
-1	0.95	0.97	0.98	0.98	0.97	0.98
-0.5	0.95	0.98	0.99	0.98	0.98	0.98
0	0.95	0.99	0.98	0.99	0.99	0.99
0.5	0.95	0.97	0.98	0.99	0.99	0.99
1	0.96	0.95	0.95	0.98	0.98	0.98
2	0.95	0.94	0.94	0.98	0.98	0.98
2.5	0.94	0.94	0.94	0.98	0.98	0.98
5	0.94	0.94	0.95	0.98	0.98	0.99
Scenario 4: Nonlinear Outcome						
-5	0.95	0.96	0.95	0.96	0.96	0.96
-2.5	0.94	0.96	0.95	0.96	0.96	0.95
-2	0.96	0.96	0.96	0.96	0.96	0.96
-1	0.96	0.97	0.97	0.95	0.96	0.95
-0.5	0.95	0.95	0.96	0.95	0.95	0.96
0	0.94	0.96	0.96	0.95	0.96	0.96
0.5	0.95	0.96	0.96	0.96	0.96	0.97
1	0.95	0.94	0.96	0.95	0.96	0.96
2	0.95	0.95	0.96	0.94	0.95	0.96
2.5	0.96	0.95	0.96	0.94	0.95	0.96
5	0.96	0.94	0.94	0.94	0.95	0.96

TABLE A4

95% coverage rates of Normal approximation confidence intervals across 1000 simulations.

Summary of Experimental Sites and Target Population

Site	Location	Expt. Size (n)	Target Pop Size (N)	Prob. of Treatment	Earnings (in \$1000) DiM	Target PATE	Employment (Percentage) DiM	Target PATE
CC	Corpus Christi, TX	524	5578	0.65	-0.21 (1.16)	1.37 (1.16)	-0.28 (3.2)	1.8 (3.2)
CI	Cedar Rapids, IA	190	5912	0.63	1.35 (1.89)	1.24 (1.89)	-0.77 (5.07)	1.71 (5.07)
CV	Coosa Valley, GA	788	5314	0.66	1.63 (0.95)	1.18 (0.95)	5.95 (2.63)	0.98 (2.63)
HF	Heartland, FL	234	5868	0.73	0.95 (1.38)	1.28 (1.38)	6.8 (5.07)	1.42 (5.07)
IN	Fort Wayne, IN	1392	4710	0.67	1.73 (0.83)	1.1 (0.83)	-0.4 (1.58)	2.23 (1.58)
JC	Jersey City, NJ	81	6021	0.64	-0.53 (3.01)	1.27 (3.01)	-2.39 (9.66)	1.67 (9.66)
JK	Jackson, MO	353	5749	0.67	2.16 (1.22)	1.19 (1.22)	5.66 (4.16)	1.38 (4.16)
LC	Larimer County, CO	485	5617	0.69	1.61 (1.32)	1.21 (1.32)	-1.97 (3.24)	1.93 (3.24)
MD	Decatur, IL	177	5925	0.70	1.24 (2.5)	1.23 (2.5)	0.03 (5.24)	1.67 (5.24)
MN	Northwest MN	179	5923	0.67	-1.43 (2.3)	1.32 (2.3)	-0.52 (6.26)	1.69 (6.26)
MT	Butte, MT	38	6064	0.71	-5.21 (4.1)	1.27 (4.1)	-7.41 (5.14)	1.67 (5.14)
NE	Omaha, NE	636	5466	0.66	1.11 (0.98)	1.25 (0.98)	-1.15 (2.56)	1.98 (2.56)
OH	Marion, OH	74	6028	0.70	-2.99 (2.71)	1.3 (2.71)	-6.82 (10.37)	1.74 (10.37)
OK	Oakland, CA	87	6015	0.64	1.83 (3.48)	1.24 (3.48)	3.34 (10.77)	1.57 (10.77)
PR	Providence, RI	463	5639	0.69	3.03 (1.34)	1.12 (1.34)	6.78 (4.58)	1.34 (4.58)
SM	Springfield, MO	401	5701	0.67	0.6 (1.31)	1.29 (1.31)	5.44 (3.34)	1.36 (3.34)

TABLE A5

Summary of the JTPA study.

Baseline Covariates included in Residualizing Models

Ethnicity		**Weeks Worked**†		**Public Assistance History**		**Family Income**†	
White		Zero		Food Stamps		Less than $3,000	
Black		1-26 weeks		Cash Welfare, other than AFDC		$3,000-$6,000	
Hispanic		27-52 weeks		Unemployment Benefits		More than $6,000	
AAPI							
		Earnings		**AFDC Histories**		**Accessibility**	
Education		Previous Earnings‡	*	Ever AFDC case head		Driver's License	
ABE/ESL		Weekly Pay	*	Case head anytime†		Car available for regular use	
High school diploma		Quantile within Site	*	Received AFDC†		Telephone at home	
GED certificate		$<$ 25%	*	Years as AFDC case head:			
Some college		$>$ 50%	*	Less than 2 years		**Household Composition**	
Occupational Training		$>$ 90%	*	2-5 years		Marital Status	
Technical certificate		Quantile across Experiment	*	More than 5 years		Spouse present	
Job search assistance		$<$ 25%	*			Household Size	
Years of Education‡	*	$>$ 50%	*	**Age**		Number of children present	
		$>$ 90%	*	Age‡	*	Child under 6 present	
Work History		Non-Zero Previous Earnings	*	Age Buckets			
Ever employed		UI Reported Earnings		20-21		**Geographic Region**	
Employed upon application				22-29		West	*
Total earnings†		**Living in Public Housing**		30-44		Midwest	*
Hourly earnings		Yes		45-54		South	*
Hours worked	*			55 or older		North	*

TABLE A6

We provide a list of all of the covariates included in the Super Learner. Many of these variables were included in the original JTPA study's regression model. Any variable denoted with an asterisk ($*$) was not included in the original JTPA study's regression model. † indicates that the measure is from the past 12 months prior to the baseline survey, ‡ indicates higher order terms included of that variable.

	Weighted Estimator			Weighted Least Squares		
	$\widehat{\tau}_W$	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\widehat{\tau}_{wLS}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$
Earnings	2.37	2.07	2.19	2.46	2.28	2.24
Employment ($\times$ 100)	8.53	8.06	8.21	7.95	7.65	8.01

TABLE A7
Mean absolute error across sites.

4.2. *Numerical Results for Empirical Application.* Table A7 provides numerical results for the mean absolute error across all 16 experimental sites for the six different estimators. We note that the mean absolute error of the point estimates do not vary substantially from using post-residualized weighting. This supports the results in Section 7.2.1.

Table A8 reports the estimated standard errors (columns 3-5 for weighted estimators and columns 8-10 for weighted least squares estimators) for each site, along with the estimated diagnostics (columns 6-7 for weighted estimators and columns 11-12 for weighted least squares estimators). In general, the diagnostics are able to adequately determine whether or not we expect there to be improvements in standard error for accounting for the population outcome information, as discussed in Section 7.2.2.

Finally, Table A9 presents the true positive rate and false positive rate for our diagnostics across the sites where the diagnostic indicated residualizing would increase precision (or not). We present these counts for both outcomes, separately.

4.3. *Using Proxy Outcomes.* To illustrate use of a proxy outcome, we run the same analysis as in Section 7, except we use employment as a proxy for earnings, and vice versa when building the residualizing model. This mimics a situation in which we have access to a related, but different outcome measure in our target population. Because employment is binary while earnings are continuous, we expect that direct residualizing may not result in substantial efficiency gains, and thus that our diagnostic measures would indicate not to residualize. However, treating $\hat{Y}_i$ as a covariate should still result in efficiency gain, as earnings and employment are correlated and the model can adjust for the scaling differences.

4.3.1. *Bias.* Table A10 presents the mean absolute error of the different estimation methods. When earnings is the outcome, both directly residualizing and using $\hat{Y}_i$ as a covariate result in relatively stable performance. However, when employment is the outcome, the scaling differences between earnings (in \$1000) and the binary employment measure lead to large residuals. We see a loss to precision from direct residualizing, and exacerbated finite sample bias. However, when including $\hat{Y}_i$ as a covariate, we are able to account for the scaling differences, and the mean absolute error is lower.

4.3.2. *Diagnostics.* We estimate the same diagnostics as in Section 7.2.1 to determine when to expect precision gains from performing post-residualized weighting. We summarize the true positive and true negative rates of the diagnostic in Table A11. We see that the performance of the diagnostic is good for direct residualization. However, we see that the diagnostic for including $\hat{Y}_i$ as a covariate is relatively conservative, and fails to identify all the cases in which it is beneficial to residualize. However, the true negative rate of the diagnostic for including $\hat{Y}_i$ as a covariate is very high (almost 100%), which indicates that the diagnostic is very effective at identifying when residualizing fails to lead to precision gain.

Table A12 provides the standard errors and diagnostic measures for each site and estimator. Within the “Weighted” and “Weighted Least Squares” sections, the left three columns present the standard error for the corresponding estimator for each site, and the right two columns present the diagnostic measure. One key takeaway is that, when employment is the

Standard Errors and Diagnostics for Residualizing Models for Residualizing Models

		Weighted					Weighted Least Squares				
Site	n	$\widehat{\tau}_W$	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\hat{R}_0^2$	$\hat{R}_{0,cov}^2$	$\widehat{\tau}_{wLS}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$	$\hat{R}_{0,wLS}^2$	$\hat{R}_{0,wLS,cov}^2$
Outcome: Earnings											
NE	636	1.70	1.53	1.53	0.23	0.22	1.58	1.53	1.53	0.08	0.06
LC	485	2.46	2.02	2.11	0.42	0.32	2.40	2.08	2.14	0.38	0.26
HF	234	1.88	1.63	1.66	0.36	0.19	1.87	1.66	1.69	0.42	0.18
IN	1392	1.03	0.93	0.92	0.25	0.26	1.00	0.91	0.91	0.22	0.21
CV	788	1.40	1.25	1.22	0.04	0.01	1.36	1.23	1.20	0.08	0.07
CC	524	2.51	2.52	2.48	-0.06	-0.18	2.42	2.42	2.39	-0.13	-0.23
JK	353	2.29	2.28	2.25	0.19	-0.25	2.19	2.18	2.16	0.30	0.10
MT	38	6.44	7.04	8.40	-0.36	-9.31	4.64	4.83	6.09	0.40	-3.90
PR	463	2.69	2.61	2.60	0.08	-0.16	2.82	2.75	2.71	0.03	-0.17
MN	179	4.79	4.70	4.80	-0.03	-0.35	3.72	4.26	4.20	-0.31	-0.56
MD	177	2.87	2.46	2.48	0.33	0.24	2.67	2.30	2.32	0.30	0.13
SM	401	2.07	2.28	2.12	-0.30	-0.13	2.13	2.23	2.11	-0.14	-0.09
OH	74	3.97	3.27	3.42	0.33	-0.22	3.94	3.75	3.77	0.29	-0.37
CI	190	3.84	3.33	3.07	0.41	0.31	3.47	3.15	2.94	0.28	-0.18
OK	87	4.69	5.07	4.64	-0.05	-0.43	4.61	4.39	4.22	0.14	-0.19
JC	81	7.24	8.81	8.50	-0.75	-1.15	6.14	7.51	6.56	-0.19	-0.49
Outcome: Employment											
NE	636	0.04	0.04	0.04	0.03	-0.01	0.04	0.04	0.04	0.02	-0.00
LC	485	0.06	0.06	0.05	0.19	0.20	0.06	0.06	0.05	0.11	0.09
HF	234	0.06	0.06	0.06	0.04	-0.03	0.06	0.06	0.06	0.02	-0.03
IN	1392	0.02	0.02	0.02	-0.15	-0.04	0.02	0.02	0.02	-0.21	-0.08
CV	788	0.03	0.03	0.03	-0.01	-0.01	0.03	0.03	0.03	-0.03	-0.01
CC	524	0.06	0.06	0.06	-0.09	-0.10	0.06	0.06	0.06	-0.10	-0.08
JK	353	0.10	0.09	0.09	0.13	-1.53	0.09	0.09	0.09	0.08	-0.85
MT	38	0.13	0.13	0.13	—	—	0.13	0.15	0.14	—	—
PR	463	0.06	0.06	0.06	0.03	-0.05	0.07	0.06	0.07	0.03	-0.03
MN	179	0.13	0.12	0.11	0.23	-3.09	0.12	0.11	0.11	0.21	-0.89
MD	177	0.09	0.08	0.08	0.19	-0.06	0.08	0.08	0.08	0.20	-4.0e28
SM	401	0.06	0.06	0.06	0.07	-0.15	0.06	0.06	0.06	0.05	-0.03
OH	74	0.07	0.06	0.07	0.09	-1.78	0.08	0.07	0.08	0.08	-1.8e28
CI	190	0.04	0.04	0.05	0.19	-0.07	0.05	0.05	0.05	0.27	-4.4e28
OK	87	0.21	0.19	0.19	0.12	-2.03	0.17	0.17	0.18	0.21	-1.35
JC	81	0.16	0.14	0.14	-0.88	-3.88	0.12	0.13	0.13	-0.92	-0.58

TABLE A8

Standard error and diagnostic values for post-residualized weighting across the 16 experimental sites for two primary outcomes–earnings and employment. The diagnostic values for the site of Butte, Montana (MT) are null when outcome is employment, because all units in the control group were unemployed.

	Weighted Estimator		Weighted Least Squares	
	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$
Earnings				
True Positive Rate	10/11	7/12	11/11	7/13
True Negative Rate	5/5	4/4	4/5	3/3
Employment				
True Positive Rate	11/13	1/12	10/10	1/7
True Negative Rate	3/3	4/4	5/6	8/9

TABLE A9

Performance of proposed diagnostic measures, as measured through the true positive rate and false positive rate.

outcome, using earnings as a proxy outcome results in large scaling differences between our residualizing model, captured by $\hat{Y}_i$, and the true outcome measure. This is unsurprising since

Estimator Performance Summary with Proxy Outcomes

	Weighted			Weighted Least Squares		
	$\widehat{\tau}_W$	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\widehat{\tau}_{wLS}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$
Earnings	2.37	2.35	2.14	2.46	2.44	2.21
Employment ($\times$ 100)	8.53	66.15	7.85	7.95	65.33	7.45

TABLE A10

Mean absolute errors for each estimator, across all experimental sites when using proxy outcomes.

	Weighted Estimator		Weighted Least Squares	
	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$
Earnings				
True Positive Rate	13/14	7/12	12/13	6/13
True Negative Rate	2/2	4/4	2/3	3/3
Employment				
True Positive Rate	–	2/12	–	3/10
True Negative Rate	16/16	4/4	16/16	5/6

TABLE A11

Performance of proposed diagnostic measures using proxy outcomes, as measured through the true positive rate and false positive rate.

earnings is continuous and employment is binary. As a result, the $\hat{R}_0^2$ measures for the estimators that use direct residualizing (i.e., $\hat{\tau}_W^{res}$ and $\hat{\tau}_{wLS}^{res}$) are all negative, indicating that we should not use direct residualizing in that setting. However, even in this scenario, the diagnostic for using $\hat{Y}_i$ as a covariate does not indicate significant gains. When using employment as a proxy for earnings, the diagnostics indicate small gains to direct residualizing across most sites, and gains from including $\hat{Y}_i$ as a covariate across about half of sites.

4.3.3. *Efficiency Gain.* Table A12 presents the standard errors of each weighting method, with and without post-residualizing, for each site. Table A13 presents the average standard error across sites for post-residualized weighting using proxy outcomes, where we restrict our attention to the sites identified by the diagnostic measures for when we expect precision gains. When using employment as a proxy for earnings, direct residualizing indicates small gains in 13/16 sites, and including $\hat{Y}_i$ as a covariate indicates gains in just under half of sites. The relative improvement in variance is small due to the differences in magnitude between $\hat{Y}_i$ and Y_i. In particular, we see around a 0.3-0.4% reduction in variance from performing direct residualizing. However, when including $\hat{Y}_i$ as a covariate, which accounts for the scaling difference, the improvements are more substantial. In particular, when using $\hat{Y}_i$ as a covariate in the weighted estimator, there is a 14% reduction in variance. Using weighted least squares, there is a 9% reduction in variance from including $\hat{Y}_i$ as a covariate. The primary takeaway to highlight is that using $\hat{Y}_i$ as a covariate to perform post-residualized weighting can allow us to leverage proxy outcomes that exist on different scales than the outcome of interest, where we expect greater gains the more closely related the outcome and proxy outcome are.

For employment, we do not consider direct residualizing because the diagnostic measure did not identify any experimental sites in which directly residualizing would lead to precision gains. When including $\hat{Y}_i$ as a covariate the diagnostic indicated 5 sites that indicate gains from post-residualized weighting; among these we see a 5% reduction in variance when using $\hat{Y}_i$ as a covariate in the weighted estimator, and a 1% reduction in variance in the weighted least squares estimator. Finally, we emphasize that estimating the PATE results in variance inflation relative to the within-sample difference-in-means, as expected. However, we see that post-residualized weighting can offset some of this loss in precision.

Standard Errors and Diagnostics for Residualizing Models with Proxy Outcomes

		Weighted					Weighted Least Squares				
Site	n	$\widehat{\tau}_W$	$\widehat{\tau}_W^{res}$	$\widehat{\tau}_W^{cov}$	$\hat{R}_0^2$	$\hat{R}_{0,cov}^2$	$\widehat{\tau}_{wLS}$	$\widehat{\tau}_{wLS}^{res}$	$\widehat{\tau}_{wLS}^{cov}$	$\hat{R}_{0,wLS}^2$	$\hat{R}_{0,wLS,cov}^2$
Outcome: Earnings											
NE	636	1.70	1.70	1.62	0.00	0.09	1.58	1.58	1.53	0.00	0.03
LC	485	2.46	2.45	2.39	0.00	0.15	2.40	2.40	2.37	0.00	0.05
HF	234	1.88	1.88	1.76	0.01	0.15	1.87	1.86	1.78	0.01	0.08
IN	1392	1.03	1.03	0.96	0.01	0.27	1.00	0.99	0.95	0.01	0.21
CV	788	1.40	1.40	1.39	0.00	-0.06	1.36	1.36	1.37	0.00	-0.03
CC	524	2.51	2.51	2.46	0.01	-0.03	2.42	2.41	2.40	0.00	-0.10
JK	353	2.29	2.28	2.10	0.01	0.07	2.19	2.18	2.07	0.01	0.04
MT	38	6.44	6.44	9.60	-0.00	-9.23	4.64	4.65	7.32	0.01	-5.86
PR	463	2.69	2.69	2.70	0.00	-0.16	2.82	2.82	2.82	-0.00	-0.15
MN	179	4.79	4.78	4.13	0.00	0.13	3.72	3.71	3.71	0.00	-0.14
MD	177	2.87	2.87	2.61	0.01	0.14	2.67	2.66	2.43	0.01	0.16
SM	401	2.07	2.07	2.04	0.00	-0.07	2.13	2.12	2.06	0.00	-0.02
OH	74	3.97	3.97	4.00	0.00	-0.44	3.94	3.93	3.75	0.00	-0.50
CI	190	3.84	3.84	3.40	0.00	-0.03	3.47	3.47	3.07	0.00	-0.22
OK	87	4.69	4.71	4.51	-0.01	-0.88	4.61	4.61	4.06	-0.01	-0.67
JC	81	7.24	7.26	8.52	-0.01	-0.82	6.14	6.17	6.75	-0.01	-0.83
Outcome: Employment											
NE	636	0.04	0.56	0.04	-352.40	-0.00	0.04	0.49	0.04	-248.43	-0.01
LC	485	0.06	0.70	0.05	-220.79	0.13	0.06	0.56	0.05	-193.43	0.02
HF	234	0.06	0.94	0.06	-260.76	-0.02	0.06	0.90	0.06	-282.80	0.06
IN	1392	0.02	0.34	0.02	-391.67	-0.04	0.02	0.32	0.02	-354.59	-0.05
CV	788	0.03	0.42	0.03	-151.95	0.02	0.03	0.38	0.03	-129.68	0.03
CC	524	0.06	1.00	0.06	-284.99	-0.21	0.06	0.89	0.06	-236.05	-0.18
JK	353	0.10	1.10	0.08	-104.99	-3.17	0.09	0.92	0.08	-88.67	-2.13
MT	38	0.13	2.42	0.12	—	—	0.13	2.49	0.13	—	—
PR	463	0.06	0.93	0.06	-228.05	-0.05	0.07	0.80	0.07	-200.67	-0.06
MN	179	0.13	1.57	0.13	-207.13	-14.37	0.12	1.61	0.12	-189.16	-3.35
MD	177	0.09	0.93	0.08	-66.00	-0.14	0.08	0.81	0.08	-77.66	-4.7e28
SM	401	0.06	0.76	0.06	-95.56	-0.12	0.06	0.68	0.06	-84.67	-0.10
OH	74	0.07	1.77	0.07	-1202.77	-0.56	0.08	1.60	0.08	-985.02	-2.4e28
CI	190	0.04	1.40	0.05	-1312.75	-0.47	0.05	1.41	0.05	-1241.70	-1.1e28
OK	87	0.21	3.24	0.16	-249.20	-1.29	0.17	2.44	0.16	-65.65	-0.28
JC	81	0.16	3.20	0.17	-6487.60	-4.51	0.12	1.70	0.14	-300.75	-0.24

TABLE A12

Standard error and diagnostic values for post-residualized weighting using proxy outcomes across the 16 experimental sites for two primary outcomes–earnings and employment. Once again, the diagnostics for MT are null when employment is the outcome, because all the units in the control group are unemployed.

This exercise shows how a proxy outcome can be used for building the residualizing model. When the two variables are on very different scales, we expect that direct residualizing would not be beneficial, as evidenced here and captured by our diagnostic measures. Including $\hat{Y}_i$ as a covariate addresses scaling concerns, although as we see when using earnings as a proxy for employment, does not always allow for gains. We see that even using proxy outcomes, our diagnostic measures can accurately capture when there is potential for precision gains, and our post-residualized weighting method can lead to precision gains in estimation of the target PATE.

REFERENCES

DING, P. (2021). Two seemingly paradoxical results in linear models: the variance inflation factor and the analysis of covariance. *Journal of Causal Inference* **9** 1–8.

JACOB, D. (2020). Cross-fitting and averaging for machine learning estimation of heterogeneous treatment effects. *arXiv preprint arXiv:2007.02852*.

Summary of Standard Errors across Experimental Sites Subset by Diagnostic, using Proxy Outcomes

	Earnings				Employment			
	Number of Sites	DiM	Standard	Post Resid. Weighting	Number of Sites	DiM	Standard	Post Resid. Weighting
Weighted								
Direct Residualizing	13	1.53	2.58	2.57	0	–	–	–
$\hat{Y}_i$ as Covariate	7	1.50	2.43	2.23	2	2.93	4.01	4.00
Weighted Least Squares								
Direct Residualizing	13	1.74	2.57	2.57	0	–	–	–
$\hat{Y}_i$ as Covariate	6	1.37	1.95	1.85	3	3.65	4.68	4.63

TABLE A13

Summary of standard errors across the 16 experimental sites identified by the diagnostic measures.

LIN, W. (2013). Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique. *Annals of Applied Statistics* **7** 295–318.

WHITE, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica: Journal of the econometric society* 1–25.